Handling Missing Data Using Worst-Case Scenario Imputation

Nieuwkasteele, Julia van

Universiteit Leiden

Faculteit der Sociale Wetenschappen

Handling missing data using worst-case scenario imputation

Julia van Nieuwkasteele

Master's Thesis Psychology,
Methodology and Statistics Unit, Institute of Psychology
Faculty of Social and Behavioral Sciences, Leiden University
Date: June 30 2021
Student number: s2398001
Supervisor: Prof. dr. M.J. de Rooij

Abstract

One universal problem in statistics is the occurence of missing data and how to handle them. In longitudinal randomized controlled trials, missing data are often observed in the form of dropout. It is important to select an appropriate technique to handle the missing data as it can lead to biased estimates. The appropriateness of techniques depends on the missingness mechanism. This thesis proposes a new developed technique called worst-case scenario imputation (WCSI) that can be used to handle MNAR data. WCSI assumes that dropout in the intervention condition is more likely for those with small to no progression, while the chance of dropout in the control condition is higher for those that do show progression. Performance of this technique was compared to maximum likelihood (ML) estimation on all available cases and multiple imputation (MI) by means of two simulation studies and one empirical study. Performance was considered under the assumption of both MNAR and MAR by investigating the bias, root mean squared error, and coverage probability associated with the parameter of interest, that is, the estimated interaction effect.

The results showed that ML and MI performed better than WCSI under the assumption of MAR, especially when only a small percentage of the dataset was missing. If the probability of dropout increased as the value of the outcome became higher in both groups (i.e. MNAR), MI and ML again resulted in the least biased estimated interaction effects. This changed when dropout depended on the true change of Y and reasons for dropout differed between the two conditions. When the dropout model was representative of the worst-case scenario, WCSI performed better in terms of bias, root mean squared error, and coverage probability. In that scenario, MI and ML overestimated the interaction effect, which could lead to wrongfully concluding that a treatment is effective. WCSI attenuated the estimated interaction effect. This was also clearly seen in the empirical study, where the originally found significant interaction effect between time and group, as found by ML and MI, was no longer present when WCSI was used.

Assuming missing data to be MAR or MNAR changes estimated interaction effects and altered the conclusions of a study. Most researchers assume MAR, while MNAR can never be ruled out. This thesis focused on a technique proposed for the situation where dropout might be responsible for finding a false effect and showed that WCSI can reduce the likelihood of a Type I error. However, when the dropout mechanism does not follow the central WCSI assumption, WCSI performs poorly. Overall, the initial results look promising, but this thesis also had its shortcomings and more research is needed. Research aimed at performance of WCSI under more complex settings is particularly interesting.

Table of contents

Section 1. Introduction

Longitudinal randomized controlled trials (RCT) are often used to investigate the effectiveness of a treatment. A randomized controlled trial (RCT) is a trial in which subjects are randomly assigned to one of two groups, one that receives the intervention being tested (intervention condition) and the other receiving an alternative treatment or no treatment at all (control condition). The two groups are followed over time to see if any differences arise in the outcome of interest. In statistical analysis of RCT data, therefore, the parameter of interest is the interaction between group and time. RCTs provide some evidence for causality as randomization ensures that there are no differences between participants in the intervention or control condition at the beginning of the study. A major advantage of longitudinal studies, compared to cross-sectional studies, is that it enables researchers to detect development or changes in the outcome of interest at both the group level and the individual level. Therefore, it can result in a better understanding of the most likely cause-and-effect relationship. Although RCTs are an effective way to analyze change, they have high costs in terms of time and money needed, and are prone to missing data (Singer & Willett, 2003; Hariton & Locascio, 2018).

Those that conduct longitudinal research are all familiar with missing data, as it is impossible to avoid when one is collecting data over time. One form of missing data often observed in longitudinal studies is dropout. Dropout refers to the situation where information is collected about a participant at the first occasion, but from a given point there is no longer information collected about this participant. In longitudinal studies, there are many possible causes for this type of missing data (Ibrahim, Chen & Lipsits, 2001). The study may take too long, the participant can move to another location or may refuse to participate due to the efficacy or even adverse effects of a study. Missing data can never be avoided completely. Often researchers assume that missing data are so common that it may not need special attention. However, if one ignores the potential influence of missing data, this may have serious consequences for the statistical power and validity of conclusions of a study (Schafer, 1997; Ibrahim & Molenbergh, 2009; Jelicic, Phelps & Lerner, 2009; Zhu, 2014).

As there can be many reasons for why data are missing in longitudinal studies, it is important to distinguish among missing data mechanisms using the appropriate terminology (Rubin, 1976; Little & Rubin, 2002). The terminology developed by Rubin in the 1970's is most widely used and distinguishes three different mechanisms: data missing completely at random (MCAR), data missing at random (MAR), and data not missing at random (MNAR). When the reasons for missingness are not related to any underlying values of the missing data

(i.e. the value that would have been observed) and is unrelated to any other variable measured, the missing data are considered to be MCAR. In this case, many complexities that arise due to missing data, with the exception of information loss, can be ignored. Data are said to be MAR when missingness is again unrelated to the value of the missing data itself, but the cause of missingness can be explained by other observed variables. One frequently mentioned example is the occurrence of missing data when analyzing weight. Women are, compared to men, more likely to refuse when asked how much they weigh, leading to missing data. In this example, the pattern of missing data will vary systematically based on gender and represents the MAR mechanism. This second category is broader and much more common than MCAR. In addition, most modern techniques to handle missing data generally start by assuming data to be MAR (Van Buuren, 2018). The last category of missing data mechanisms is the mechanism of data that are MNAR. In this last case, the missingness is related to the underlying value of the missing data itself. For instance, participants that weigh too much or too little are more likely to not answer questions about their weight than participants with average weight.

There are different techniques to handle missing data and appropriateness of each technique highly depends on the missing data mechanism, therefore, it is important to identify the missing data mechanism before choosing a technique. In longitudinal trials it is difficult to test the underlying mechanism. One can check for systematic differences for participants with and without missing responses using a set of observed variables to see whether data are MAR or MCAR. However, in real application, data being MCAR are least likely and often means that the variables that could explain systematic differences have not been included in the study. More important is to keep in mind that it is not possible to test whether data are MNAR as one does not know the true values of the missing data. A particular condition of missing data mechanism can therefore only be stated as an assumption and not as a truth (Musil, Warner, Yobas & Jones, 2002; Jelicic, Phelps & Lerner, 2009).

Although the missing data mechanism can only be assumed, it is necessary that one accounts for the missing data in data analysis in order to draw valid conclusions. Most recently developed techniques assume MAR data and overcome the serious violations of statistical assumptions observed when missing data was ignored. Some of the techniques used in the past, such as substituting missing values with the mean or single imputation, have been shown to be potentially highly misleading (Allison, 2002; Graham et al., 2003; Peugh & Enders, 2004; Shafer & Graham, 2002).

Some techniques developed more recently, such as multiple imputation (MI) and maximum likelihood estimation (ML) are based on both a theoretical framework and statistical theory concerning missing data and lead to less biased and more accurate conclusions, especially when the sample size is large ($n > 1000$) and the amount of missing data is small (Black, Harel & McCoach, 2010; Cheema, 2014). In the presence of MAR, these methods can give unbiased results. The assumption of MAR may not always be plausible, in particular with clinical trials (Sterne et al, 2009), and often sensitivity analyses are needed to get a better idea of the influence of MNAR data on the estimated effects. By replacing all missing values with the worst/best outcome, one can get a better idea of the full range of potential outcomes that must be taken into account before making statements (Little et al., 2012; Morris, Kahan & White, 2014).

MNAR can be seen as the most complex case as it is non-ignorable. This means that the missing data mechanism itself has to be modeled when one deals with missing data to obtain correct inferences. Model-based approaches can be used to deal with MNAR data (Verbeke & Molenberghs, 2014; Chen et al., 2018; Fiero, Hsu & Bell, 2018). Two main approaches that have been proposed to handle longitudinal MNAR data are selection models (Little & Rubin, 2002) and pattern mixture models (Little, 1993). Both approaches rely on strong assumptions that can never be proven with observed data only, heavily depend on expert opinions about the possible range of potential influences, and are often advised to be part of sensitivity analyses as well. Both approaches have been criticised and more research is needed to develop techniques to handle data that are MNAR, especially techniques that rely on more plausible assumptions (Michiels, Molenberghs, Bijnens, Vangeneugden &, Thijs, 2002; Kaciroti & Raghunathan, 2014; Van Buuren, 2018).

The aim of this thesis is to propose a new missing data technique called worst-case scenario imputation (WCSI). WCSI assumes that, in randomized controlled trials, the likelihood of dropout differs between the two conditions. In the intervention condition dropout is assumed to be more likely for those that show small or no progression, while it is assumed that those in the control condition are more likely to show dropout when they do show progression. This thesis will perform a first investigation into the accuracy of multilevel estimates under the assumption of MNAR and MAR using WCSI. The idea is that WCSI can lead to unbiased estimates under the assumption of MNAR, especially when missing data are assumed to represent the worst-case scenario. This thesis is organized as follows. The next section, section 2, will provide more information about MI and ML, and introduces the technique of worst-case scenario imputation and its relevance. In the third section, the

performance of the three missing data techniques will be assessed by means of two simulation studies. Section 4 will focus on the performance of these techniques by using one empirical dataset from a study by Van Luenen, Garnefski, Spinhoven and Kraaij (2018). The last section includes a summary and discussion of all results, limitations of this thesis and some suggestions for further research.

Section 2. Missing data techniques (MDTs)

There are multiple techniques that can be used to handle missing data. Many of these techniques have been developed within the field of longitudinal clinical trials, but often researchers still use inappropriate methods (Roth, 1994; Peugh & Enders, 2004). Compared to older, less appropriate techniques, maximum likelihood estimation and multiple imputation have been established as good alternatives, especially under the assumption of MAR (Newman, 2003).

2.1 Maximum likelihood estimation

The MAR-based maximum likelihood estimation method can be used in data analysis, such as multilevel modelling, on incomplete datasets. It defines a model based on the observed data and makes inferences using likelihood functions, meaning that the computation process operates as if missing data are replaced with values most likely given the observed values of other variables by using linear relationships between the variable with missing data and other variables in the model (Enders, 2011). These values for the missing data are only implied to obtain the final estimates and are not imputed within the dataset. It is currently seen as an appropriate and easy to implement technique to handle MAR data. Multilevel models can be easily fitted using maximum likelihood methods to take care of missing data, if missingness occurs in only the dependent continuous variable. Doing so leads to more accurate estimates compared to traditional methods such as single mean imputation. This method does assume data to be MAR in order to lead to reliable results and is especially advised when the analytical model is complicated (e.g. with interactions) to avoid any problems of model compatibility between the analytical and imputation model (Enders & Bandalos, 2001; Jakobsen & Gluud, 2017; Van Buuren, 2018; Chen, Li & Liu, 2018).

2.2 Multiple imputation

In contrast to maximum likelihood estimation, multiple imputation solves the problems associated with incomplete datasets by producing more than one imputed dataset. Rubin (1976; 1987) proposed multiple imputation as a way to analyze incomplete data under the missing mechanism of MAR. As multiple imputation replaces the missing values with estimates, one of the advantages of this technique is that it enables the use of complete-data methods for analysis. These missing values are replaced more than once and each imputed dataset contains slightly different values. Rubin (1987) claims that good inferences can be made when 3 to 5 datasets are imputed. Others suggest that the number of imputed datasets,

m, should be based on the amount of missing data and the amount of difference in estimations one is willing to allow between imputed datasets (Van Buuren, 2018; Von Hippel, 2018).

All standard procedures, that are developed and normally used for the analysis of complete datasets, can be applied to each of the *m* datasets. These procedures are thus conducted on multiple datasets, and the results from each of these analyses are combined using Rubin's pooling rules. This way a single parameter estimate (i.e. the average of multiple parameter estimates from the imputed datasets) and its standard error (i.e. function of both the average of multiple standard errors from each imputed dataset and an added term that captures variability in the estimates across imputations) are computed. MI is superior to single imputation methods by introducing variability and taking into consideration the uncertainty in estimates that occur when imputation is used. One downside of this technique is that it requires more effort, time and computer storage compared to simpler techniques such as maximum likelihood estimation. At this point in time, MI methods for missing data are one of the best performing in terms of giving most valid results and the use of MI increased over the years (Rubin, 1987; Rezvan, Lee & Simpson, 2015; Van Buuren, 2018).

MI can also be used to impute multilevel data and over the years different ways to impute this type of data have become available. Selection of the best method depends on the level of the variable to be imputed. In this thesis, the missing data was always generated to be or observed in the continuous level-1 dependent variable. One method suggested to impute missing values in such a continuous variable is the pan method. This multilevel method uses a linear two-level model with homogeneous variances to draw univariate imputation using a Gibbs sampling procedure. One major advantage of this method is that it allows for specification of different roles for the predictor variables included in the imputation model (e.g. the model can include both random and fixed effects) and is recommended for the imputation of multilevel data (Schafer & Yucel, 2002; Grund, Ludtke & Robitzsch, 2016).

2.3 Worst-case scenario imputation

In real life situations, it is impossible to exclude the possibility of data being MNAR completely. Two of the best techniques to handle missing data (ML and MI) assume data to be at least MAR. This may be a good starting point, however it may not be realistic considering the data itself. When the data are not MAR, one strategy often used is to make the data "more MAR" by identifying additional sources that help explain why data are missing. This additional information can then be used to generate imputations conditional on that information. Another strategy is to perform sensitivity analysis, where imputations are

generated according to certain scenarios. One frequently used scenario is the unrealistic worst-case scenario, this means that one assumes that all participants that show dropout have scored the highest or lowest possible value. Such extreme scenarios are highly unlikely but can be used to explore the influence of assuming the worst and best on model estimates (Little et al., 2012; Morris, Kahan & White, 2014; Van Buuren, 2018).

For this thesis, the aim is to propose a more sophisticated method that can be used to impute values when one assumes that dropout represents a more likely and specific worst-case scenario. Instead of assuming MAR, it is assumed that data are MNAR. In the intervention condition, dropout is assumed to be more likely for participants that show small to no progression (i.e. there is no to small relief of symptoms or even an increase of symptoms). This can be the case when dropout occurs due to adverse effects of treatment or is ineffective. On the other hand, in the control condition dropout is assumed to be more likely for participants that do show progression (i.e. there is a relief of symptoms) which could be due to a placebo effect or natural recovery. The aim of WCSI is to account for this pattern during imputation. In the worst-case scenario data are MNAR and assuming that data are MAR would lead to biased results. More specifically, the likelihood of dropout depends on the value of outcome itself and differs based on the condition.

WCSI is highly similar to multiple imputation as the missing values will be imputed more than once and the function developed to perform WCSI also uses the pan method discussed before. The difference between WCSI and MI lies in the amount and specific part of the data used for imputation. For WCSI, only a specific part of the dataset is selected by calculating the change scores of participants from two time points and selecting only a percentage of the dataset that has shown the most or least progression over time. Information from the selected dataset of participants in the intervention condition with most progression is then used to impute the missing data observed in the control condition *m* times. This process is repeated exactly the other way around for participants in the intervention condition. Imputation of their missing data is based only on information from the participants in the control condition with the least progression. This way the *m* imputed values will capture the worst-case scenario. Values imputed for those in the intervention condition are now more likely to represent the least progression, while values imputed in the control condition are more likely to represent a form of progression. As the imputed values are still based on true data, and not selected to be the worst/best value possible, it is assumed that this technique will result in realistic estimates under the assumption of MNAR and takes the worst-case scenario into account.

As the presence of MNAR can never be truly excluded, research into techniques for handling MNAR data is still active, and sensitivity analyses only give some information about the possible range of values one should take into account, it is of great interest to investigate the accuracy of this new technique. In addition, this technique focuses on the unfavorable situation in research where the possibility of making Type I errors has increased due to dropout. In the worst-case scenario, there is an increased chance of incorrectly concluding that the treatment has the intended effect when in fact there is none. Loss of participants with intended results in the control condition and dropout in the intervention group for those without any progression can lead to a distortion of the overall effects. The treatment will appear to be more favorable than it actually is. When it is incorrectly concluded that the treatment is effective, a lot of effort, time and money can be lost during implementation before the absence of intended treatment effects is discovered.

This thesis will show what happens to the estimated interaction effects in a multilevel model when missing data are handled using worst-case scenario imputation in comparison to well-established methods such as maximum likelihood estimation and multiple imputation under both MAR and MNAR. The expectation is that ML and MI will be more accurate when data are MAR but will perform less than WCSI when data are MNAR and the mechanism represents the worst-case scenario explained before. In addition, WCSI is less likely to result in estimates that strongly represent effective treatment in comparison to ML and MI.

Section 3. Simulation studies

In this section, the methods used and results found for two simulation studies will be presented. Both simulation studies assessed the performance of the three aforementioned techniques to handle missing data using longitudinal multilevel data. The first simulation study replicates the study of Kenward and Diggle (1994) by generating MAR and MNAR data. The second study replicates the effects found by Van Luenen et al. (2018) and generates MNAR data. In the second simulation study, the influence of using different percentages of the dataset in WCSI will be assessed to a larger extent compared to the first simulation study. The main difference between the simulation studies lies in the model used to generate MNAR data. In the second simulation, the worst-case scenario is clearly represented by the dropout model. Whereas, in the first simulation study the MNAR condition does not follow the WCSI assumption.

3.1 Simulation study (Kenward and Diggle (1994))

3.1.1 Data generation

In this simulation study, longitudinal data was generated based on the following multilevel model,

$$y_{ij} = \gamma_{00} + \gamma_{01}\,\text{group} + \gamma_{10}\,\text{time} + \gamma_{11}\,\text{group} * \text{time} + u_{0j} + u_{1j}\,\text{time} + \varepsilon_{ij}, \qquad (1)$$

where $\varepsilon_{ij} \sim N(0, \sigma^2_e)$, and

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}\right).$$

In Equation 1, y_{ij} is the outcome of individual *i* at occasion *j*, time and group are explanatory variables. The first part, $\gamma_{00} + \gamma_{01}$ group $+ \gamma_{10}$ time $+ \gamma_{11}$ group * time, refers to the fixed effects of the model. The last term, γ_{11} group * time, is a cross-level interaction, which signals that the slope varies with the group level variable. The coefficients of the fixed effects represent the average within-group population intercept and slope.

The last part of the equation, $u_{0j} + u_{1j}$ time $+ \varepsilon_{ij}$, represents the random effects of this model. These random effects, u_{0j} and u_{1j}, are assumed to be randomly drawn from a multivariate normal distribution with mean structure (0, 0) and a variance-covariance matrix.

For this simulation study, random effects are included for the intercept and time resulting in a 2x2 matrix with the following components:

- σ^2_0 is the variance of the random intercepts;
- σ^2_1 is the variance of the random slopes;
- σ_{01} is the covariance between the within-individual intercepts and slopes.

The last term of the equation, ε_{ij}, refers to the residual errors for each individual. These are drawn from a normal distribution ~ $N(0, 1)$.

Following the condition of Kenward and Diggle (1994), a randomized controlled trial was simulated using 1000 datasets under the following conditions: (i) the number of groups was set to two, (ii) the number of participants was set to 50 with a probability of being assigned to either group of .5, and (iii) there were ten equally spaced measurements over time. The value of the intercept, γ_{00} , was fixed at 10 for both groups, as the effect of group, γ_{01}, was set to zero. The effect of time, γ_{10}, was also set to zero. However, the cross-level interaction effect between time and group, γ_{11}, was set to -1. The random components of this model, u_{0j} and u_{1j}, are assumed to drawn from a multivariate normal distribution with mean zero and standard deviations σ_0 andσ_1. For this simulation study, the standard deviation of the intercept, σ_0, was set to .75 and the standard deviation of the slope, σ_1, was set to .2. The covariance was set equal to .1, so that the correlation between intercept and slope was a moderate to strong correlation of .67.

3.1.2 Dropout model

After creating the complete datasets, some of the observed values were set to be missing, based on either the MNAR or MAR missing mechanism. Missing data was generated based on the following drop-out model,

$$\text{logit}(p_{\text{missing}}) = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1}.$$

Here the logit of the probability of an observation being assigned a missing value depends on a constant, β_0, β_1 multiplied by the value of the outcome of interest at time point *t* and β_2 multiplied by the value of the outcome of interest at a previous time point. Under both MNAR and MAR mechanisms, β_0 was set to a small value.

In the MAR setting, the probability of being a missing value did not depend on the observation of Y itself. It did depend on the value of Y at a previous time point. Therefore, β_1 was set to zero, while β_2 was set to values that would result in approximately 10 or 25 percent missing data. The reverse is true in the MNAR setting, where the probability of being missing did depend on the observed value of Y itself and there is no dependence on previous observations of Y. This time, β_2 was set to zero in order to eliminate the relationship with measurements of Y at previous time points and β_1 was set to generate approximately 10 or 25 percent missing data.

To determine the values of β needed to simulate the preferred amount of missing data, a large dataset of 10.000 observations was generated using the design as specified before. Under both MNAR and MAR mechanisms, approximately 10 percent missing data was simulated by setting β_0 to -5, and β_1 and β_2 to .135. To simulate approximately 25 percent missing data, β_0 was again equal to -5, however β_1 was set to .26 and β_2 equal to .25.

In this simulation study, two factors were varied: (i) missing data mechanism used to generate missing data (MAR and MNAR), and (ii) the percentage of missing data (approximately 10 and 25 percent). Missing data always followed a monotone pattern, meaning that once a missing value was observed for a participant, all subsequent measures also showed missing values.

3.1.3 Procedures

All multilevel models were fit on all datasets using the lme() function from the Linear and Nonlinear Mixed Effects Models (nlme) package (Pinheiro, Bates, DebRoy & Sarkar, 2021). Time and group were included as fixed effects and slopes for time and the intercept as random effects. The optimizer was set from the default (nlminb) to optim and the number of iterations was set to 500 to increase the likelihood of convergence. Convergence of a model occurs when the estimation procedure stabilizes upon a unique solution, problems may arise when missing data results leads to insufficient information about the parameters or random effects have small variances (Black, Harel & McCoach, 2011). The estimation method was set to maximum likelihood estimation. First, the multilevel model was fitted on the incomplete datasets to represent handling missing values by maximum likelihood estimation as suggested by Ibrahim and Molenbergh (2009).

Secondly, multiple imputation was performed using the mice() function from the mice package (Van Buuren & Groothuis-Oudshoorn, 2011). The imputation method selected was pan (Schafer & Yucel, 2002). Each missing value was imputed 5 times as suggested by Rubin

(1987). For the imputation model, time as a numerical value was included as a predictor with both a fixed and random slope effect. The group variable and dummy variables for the interaction were included as fixed effects. Id was specified as the cluster variable. The simulated datasets did not include any other variables.

Finally, worst-case scenario imputation was performed using the R function found in appendix B. All settings, such as number of imputations and specification of predictor variables, mentioned for MI apply here with the exclusion of the group variable as a predictor. This technique was used twice with different amounts of the data used for imputations. The first time 10 percent of the other condition with the most or least progression was selected (WCSI 10%), while the second time 25 percent of the dataset was used (WCSI 25%). The change scores used were calculated by taking the difference of the outcome from the beginning to the end of the study.

3.1.4 Performance measurements

Bias, root mean squared error (RMSE) and coverage probability will be used to assess the accuracy of each missing data technique. The main focus will lie on the difference between groups over time captured by the interaction effect. The accuracy of this parameter estimate can be quantified by the bias. Let θ stand for the true population parameter, then

$$\sum_{r=1}^{R} \frac{(\hat{\theta}_r - \theta)}{R},$$

where $\widehat{\theta}_r$ is the parameter estimated for the *r*th replication and R is the number of simulated datasets.

The mean squared error (MSE) is defined as the average squared difference between the estimated parameter ($\widehat{\theta}_r$) and the corresponding true parameter (θ). RMSE is the square root of the MSE and can be seen as a measure of overall precision. This can be calculated by

$$\sqrt{\sum_{r=1}^{R} \frac{(\hat{\theta}_r - \theta)^2}{R}}, .$$

In general, more effective techniques would have smaller bias and lower RMSE values.

The last performance measurement is the coverage probability (CP), which is defined as the proportion of simulated datasets, among the total amount of simulated datasets, of which the constructed 95 percent confidence interval contains the true parameter. The 95 percent confidence interval is defined as

$$\hat{\theta}_r \pm t_{df,\ 1\ -\ \alpha/2} * SE_r,$$

where $\hat{\theta}_r$ is the estimated interaction effect, *t* is the t-statistic, df is the degrees of freedom and SE_r is the associated standard error; MI and WCSI use the pooled estimated effects and pooled standard error. An appropriate method should have a coverage probability of around 95 percent.

3.1.5 Results

This section first discusses the performance of each missing data technique on 1000 datasets with approximately 10 percent missing data; an overview of the results is given in Table 1.

Table 1

Overview of results with approximately 10 percent missing data (r = 1000)

Missing data technique	Missing data mechanism	Bias	RMSE	CP in %
ML	MAR	0.0021	0.0675	93.8
MI	MAR	0.0016	0.0684	95.8
WCSI (10%)	MAR	0.0658	0.1122	98.2
WCSI (25%)	MAR	0.1013	0.1368	92.3
ML	MNAR	0.0033	0.0657	94.5
MI	MNAR	0.0030	0.0675	96.7
WCSI (10%)	MNAR	0.0700	0.1144	97.8
WCSI (25%)	MNAR	0.1041	0.1374	93.8

For each of the three different missing data techniques, performance was assessed using the bias, RMSE and coverage probability. The method of worst-case scenario imputation was used twice. Once 10 percent of the information from the other condition was used for imputation and a second time 25 percent of the dataset was used.

Under the MAR mechanism, where missingness depended on the observation of Y at the previous time point, both the use of maximum likelihood estimation and multiple imputation resulted in virtually unbiased estimates of the interaction effect between time and group. The bias that results from using maximum likelihood is equal to .0021, which is only slightly larger than the bias associated with multiple imputation of .0016. Using worst-case scenario imputation does not lead to unbiased estimates of the interaction effect as indicated by the calculated bias.

When 10 percent of the data was selected, the associated bias is .066 and bias is even as high as .101 when more information is used for imputation (WCSI 25%). Investigation of the calculated RMSE for each technique shows that the estimated effects are least spread out when ML or MI was used. ML and MI perform the best with RMSE as small as .07, while WCSI with 10 percent results in a RMSE value of .11 and using more information (25%) increases the value to .14. The coverage probability is closest to the preferred value of 95 percent when multiple imputation is used, followed by the coverage probability associated with maximum likelihood. Using worst-case scenario imputation with 10 percent of the data gives a coverage probability higher than 95 percent. Although the coverage probability of WCSI (25%) was lowest of all with a value of 92.3. All techniques resulted in a coverage probability close to the preferred value of 95 percent.

Results are quite similar when data are missing under the mechanism of MNAR, where the probability of missingness depends on the value of Y itself. Higher values of Y are more likely to result in dropout for participants in both conditions. Based on bias, RMSE and CP, multiple imputation and maximum likelihood estimation on all available cases perform best. With bias around .003, these techniques for handling missing data result in the least biased estimates for the interaction effect when dropout is simulated to be MNAR. The calculated values for RMSE of .07 indicate that these techniques have more similar estimated effects compared to WCSI. Estimated values for the interaction effect become more biased when worst-case scenario imputation is used and the bias tends to grow larger as more cases are used to impute the missing data. Maximum likelihood results in a coverage probability closest to 95 percent, but again all calculated probabilities are close to 95. When approximately 10 percent of the data are missing either MAR or MNAR, maximum likelihood estimation and multiple imputation perform better compared to worst-case scenario imputation.

Table 2

Overview of results with approximately 25 percent missing data (r = 1000)

Missing data technique	Missing data mechanism	Bias	RMSE	CP in %
ML	MAR	-0.0012	0.0764	95.1
MI	MAR	-0.0044	0.0813	97.4
WCSI (10%)	MAR	0.0890	0.1565	98.0
WCSI (25%)	MAR	0.1601	0.2030	92.4
ML	MNAR	0.0111	0.0785	92.8
MI	MNAR	0.0074	0.0830	97.2
WCSI (10%)	MNAR	0.1029	0.1609	97.2
WCSI (25%)	MNAR	0.1710	0.2068	91.3

An overview of the performance of each technique on datasets with approximately 25 percent missing data can be found in Table 2. These results show that even when more data are missing, using maximum likelihood estimation and multiple imputation still leads to virtually unbiased estimated effects of the interaction effect in the case of data MAR. As before, the estimated effects become further from the true parameter when worst-case scenario imputation was used and this discrepancy grows larger when a higher percentage is used for imputation. It is interesting to see that ML and MI result in negative bias in the situation where MAR was manipulated. This means that on average the estimated effects are estimated to be stronger (i.e. estimated effects lower than -1.15). Worst-case scenario imputation resulted in underestimates of the true parameter (i.e. estimated effects higher than -1.15). Coverage probability was close to 95 percent for each missing data technique. Maximum likelihood estimation seems to be the best fit here based on all measurements.

In the case of data that is MNAR, multiple imputation should be preferred over the other techniques based on the bias. Maximum likelihood estimation also results in unbiased estimates of the true interaction effect between time and group. Looking at the associated RMSE and coverage probabilities, ML and MI have similar values. Worst-case scenario imputation leads to estimated effects most biased in comparison to the other two techniques and as before the error grows larger with the inclusion of more information.

3.2 Simulation study 2 (Van Luenen et al. (2018))

3.2.1 Data generation

Longitudinal data was generated using the same multilevel linear model as in the first simulation study, see Equation 1. For this simulation study, a fixed covariate X for each unit was included. This covariate X was randomly drawn from a normal distribution, $X_i \sim N(0, \sigma^2)$.

Following the results from the study of Van Luenen et al. (2018), using maximum likelihood estimation on all available cases, 1000 datasets were simulated with the following conditions: (i) number of groups was set to two, (ii) sample size was set to 200 with a probability of .5 to be assigned to one of the conditions, and (iii) three equally spaced measurements over time for each participant. For this study, the intercept, γ_{00} , was fixed at 12.4. Contrary to the estimates from the empirical study, the effect of group, γ_{01}, was set to zero to model the ideal situation where the conditions do not differ at baseline. Time, γ_{10}, had an overall effect of -1.57. The estimate of interest was again the cross-level interaction effect between time and group, γ_{11}, which was set to either -1.15 or 0. As before, the random effects of the model, u_{0j} and u_{1j}, are assumed to drawn from a multivariate normal distribution with mean zero and standard deviations σ_0 and σ_1. The standard deviation of the intercept, σ_0, was set to 2.44. and the standard deviation of the slope, σ_1, was set to .57. The correlation between intercept and slope was set to .56, resulting in a covariance of .78 The standard deviation of the error term was set to 3.6 and covariate was drawn from a distribution with mean 0 and standard deviation of 5.

One of the two factors that varied in this simulation study is the true interaction effect between time and group. The true parameter was set to -1.15 or to zero (i.e. absence of interaction effect). The second factor concerns the amount of missing data (approximately 10 or 25 percent), more details are given in the next section.

3.2.2 Dropout model

In this study, the missing values produced represent the worst-case scenario. This means that the likelihood of being a missing value (indirectly) depends on the value of Y itself, which is known as MNAR. More precisely, missingness will depend on the true change in Y for each participant in this simulation study. True change in Y is captured by the random slope effects, u_{1j}.

For participants in the intervention condition, dropout is assumed and modelled to be more likely when there is a lack of progression or even adverse treatment effects (i.e. an increase in Y) from the first measurement to the last one. This means that the effect of treatment is not as desired. The dropout model also included a small effect based on the covariate, where higher values of X increase the likelihood of a missing value. The dropout model is

$$\text{Logit}(\text{Probability of } Y_{ij} \text{ being missing} \mid \text{Group is Intervention}) = \beta_0 + \beta_1 u_{1j} + \beta_2 X.$$

Here the logit of the probability of an observation being assigned a missing value depends on a constant, β_0, β_1 multiplied by the value of the true change, and β_2 multiplied by the value of the covariate X.

In the control condition, the likelihood of a missing value increases when there is a negative true change (i.e. progression in Y during the study). This means that dropout in the control group is more likely for participants that show the effect that is typically observed in the intervention. Here the covariate X has the same effect on the likelihood of dropout (i.e. higher values increases the likelihood of a missing value). For the control condition, the dropout model is specified as

$$\text{Logit}(\text{Probability of } Y_{ij} \text{ is missing} \mid \text{Group is Control}) = \beta_0 - \beta_1 u_{1j} + \beta_2 X.$$

Here the logit of the probability of an observation being assigned a missing value again depends on a constant, β_0, β_1 multiplied by the value of the true change, and β_2 multiplied by the value of the covariate X. For both models, β_0 was set to a small value to ensure that the influence of random factors on drop-out was negligible.

To determine what values are needed to generate approximately 10 and 25 percent missing data, a large dataset of 3000 observations was generated (N = 1000) following the design as specified before. β_0 was always set to -5, while β_2 was equal to .08 in all conditions. To generate approximately 10 percent missing data, β_1 was set to 8, and β_1 was set to 20 to generate approximately 25 percent missing data.

3.2.3 Procedure

Procedures used to fit the multilevel models on the datasets are similar to the first simulation study. The multilevel model included time and group as fixed effects and slopes for time and the intercept as random effects. For MI and WCSI, this simulation study included the same predictors in the imputation model as in the first simulation, but also covariate X was included as fixed effect. In this simulation study, worst-case scenario imputation was not used only twice (i.e. 10 and 25 percent of the data), but used five times. Imputations were generated using 10 percent (WCSI 10%), 25 percent (WCSI 25%), 50 percent (WCSI 50%), 75 percent (WCSI 75%) and 100 percent (WCSI 100%). In the last situation, this means that all complete observations from one condition were used to impute the missing values in the other condition.

3.2.4 Performance measurements

In this simulation study, performance of each missing data technique will again be assessed by calculating the bias, root mean squared error (RMSE) and the 95 percent coverage probability rates. See section 3.1.5 for more details. The main focus again lies on the difference between groups over time captured by the estimated interaction effect between time and group.

3.2.5 Results

First, the results will be discussed when the true interaction effect used to generate the data is set to -1.15. The second part of the results will focus on the estimates when data are generated without a true interaction effect ($\gamma_{11} = 0$).

Table 3 gives an overview of results of the 1000 datasets where approximately 10 percent of the data was missing. When data are MNAR, based on the true change and representative of the worst-case scenario outlined before, fitting a multilevel model using maximum likelihood on the incomplete dataset results in similar bias, RMSE and CP as fitting a multilevel model after multiple imputation. Both methods show bias around -.65 and RMSE of .76, and their coverage probabilities of 65 and 67 percent are close too. Interesting is that the bias is positive when worst-case scenario imputation was used. This shows that the average estimated effect is less strong than the true effect, meaning that the interaction effect between group and time has been toned down by WCSI. As the amount of data used for imputation increases from 10 percent to 100 percent, the bias and RMSE become smaller, while the coverage probability increases. When 50 percent of the dataset is used, performance

of WCSI is similar to that of ML and MI. Although bias is of similar size, ML and MI seem to result in overestimations of the interaction effect, while WCSI attenuated this effect. In the situation where approximately 10 percent of the data are MNAR based on the true change, using all data from the other condition to impute missing values on average leads to estimated interaction effects closest to the true parameter as indicated by the bias of .043 and RMSE of .31. The coverage probability of 99.5 percent is quite high and shows that almost all 95 confidence intervals contain the true parameter, this signals wide confidence intervals due to larger standard errors.

Table 3

Overview of results with approximately 10 percent missing data (γ_{11} = -1.15 / *r* = 1000)

Missing data technique	Bias	RMSE	CP in %
ML	-0.6489	0.7640	64.8
MI	-0.6477	0.7671	67.2
WCSI (10%)	0.9859	1.0800	60.4
WCSI (25%)	0.9105	0.9820	53.8
WCSI (50%)	0.6436	0.7214	73.0
WCSI (75%)	0.3685	0.4795	92.3
WCSI (100%)	0.0432	0.3100	99.5

The same multilevel models were again fit on 1000 datasets with approximately 25 percent missing data using different techniques to account for MNAR data (see Table 4).

Table 4

Overview of result with approximately 25 percent missing data (γ_{11} = -1.15 / *r* = 1000)

Missing data technique	Bias	RMSE	CP in %
ML	-1.2631	1.3360	16.0
MI	-1.2597	1.3382	24.9
WCSI (10%)	2.0681	2.1623	34.2
WCSI (25%)	1.9468	1.9976	9.5
WCSI (50%)	1.4745	1.5151	11.5
WCSI (75%)	0.9944	1.0383	42.7
WCSI (100%)	0.4499	0.5321	95.3

The overall pattern observed in datasets with 25 percent missing data is very similar to the pattern discussed before. Maximum likelihood estimation and multiple imputation lead to equivalent bias around -1.26 and RMSE of 1.3. Once again the coverage probability of multiple imputation is a little higher, although both missing data techniques perform poorly in this scenario where data are MNAR and representative of the worst-case scenario based on true change. In this simulation study, WCSI performs even worse in terms of bias, RMSE and coverage probability. This only changes when 75 or 100 percent of the dataset is used for imputation. With 75 percent of the dataset used for imputation, the bias and RMSE are around 1 and the CP is above 40 percent. When 100 percent of the other group is used for imputation, this leads to the least biased estimated effect as indicated by bias of .45, RMSE of .53 and a coverage probability of 95 percent. One interesting observation that can be made under this design is that WCSI always resulted in positive bias, while the other techniques resulted in negative bias. ML and MI on average lead to estimated interactions effects stronger than the true parameter, while WCSI lead to estimated interactions effects that are weaker or even in the opposite direction when the true estimated interaction effect is set to -1.15.

Table 5 gives a summary of the accuracy of each missing data technique in the situation where the true parameter for the interaction effect between group and time has been set to zero and approximately 10 percent of the data was manipulated to show missing values. As before, maximum likelihood estimation and multiple imputation provided similar results based on bias, root mean squared error and coverage probability. Their associated bias was again negative, meaning that the estimated effects would signal a decline in symptoms within these simulated datasets that represent randomized trials. Worst-case scenario imputation using 10, 25 and even 50 percent of the data from the other condition for imputation of missing values resulted in positive bias, meaning that the estimated effects on average would have shown an increase in symptoms. The estimated effects for the interaction between time and group become negative when 75 percent or more is used for WCSI. In this simulation study, use of 75 percent of the data with WCSI gives the most accurate results in terms of bias, RMSE and CP. In the absence of a true interaction effect, WCSI can lead to over- or underestimations of the effect. The estimates are least biased when 75 percent is used, on average WCSI results in an estimated interaction effect of -0.073 (RMSE = 0.319, CP = 98.3%). WCSI always results in estimated effects that signal the absence of an interaction effect more closely than ML and MI.

Table 5

Overview of results with approximately 10 percent missing data (γ_{11} = 0 / r = 1000)

Missing data technique	Bias	RMSE	CP in %
ML	-0.6757	0.7897	61.3
MI	-0.6760	0.7936	64.7
WCSI (10%)	0.6036	0.74023	83.4
WCSI (25%)	0.4914	0.6083	85.9
WCSI (50%)	0.2031	0.3767	98.0
WCSI (75%)	-0.0734	0.3186	98.3
WCSI (100%)	-0.4061	0.5127	91.1

When more data are generated to be missing, the estimated effects of the interaction have become more biased. This can be seen in Table 6 below, where an overview is given of the performance of all techniques using 1000 datasets in which the true parameter was set to zero and approximately 25 percent of the data was omitted.

Table 6

Overview of results with approximately 25 percent missing data (γ_{11} = 0 / r = 1000)

Missing data technique	Bias	RMSE	CP in %
ML	-1.2109	1.2882	21.9
MI	-1.2009	1.2809	28.2
WCSI (10%)	1.3701	1.4940	62.2
WCSI (25%)	1.1567	1.2276	48.2
WCSI (50%)	0.6741	0.7442	81.9
WCSI (75%)	0.1922	0.3363	98.9
WCSI (100%)	-0.3483	0.4465	97.6

The estimated effects that result from using maximum likelihood estimation and multiple imputation are on average as large as -1.2 with RMSE around 1.3 and coverage probabilities below 30 percent. These estimated effects are even further from the truth when worst-case scenario imputation with 10 percent of the dataset is used. This is indicated by bias of 1.4 and RMSE of 1.5, although the coverage probability of 62.2 is higher compared to the other techniques. WCSI using 10, 25, 50 and even 75 percent of the dataset leads to estimated effects that are positive (i.e. signal an increase in Y). Using 100 percent did again result in a negative estimated effect, which is much smaller than the negative effects estimated using ML and MI to handle missing data. In this scenario, using WCSI with at least 25 percent of the data will always give estimated effects closer to the absence of an effect compared to ML and MI. This is also indicated by the higher values of the associated coverage probabilities. Using WCSI with 75 percent resulted in estimated effects that were most accurate. As before, WCSI performs best when a larger percentage (i.e. minimum of 75 percent) is used.

Section 4. Empirical study

4.1 Dataset

Data used to assess the influence of different missing data techniques on the direction and strength of the estimated interaction effects was collected as part of a randomized controlled trial by Van Luenen et al. (2018). For their study, 188 participants above 18 years old were recruited from 23 HIV treatment centers. All participants had been diagnosed with HIV at least six months before the study and had mild to moderate depressive symptoms as assessed by the Patient Health Questionnaire-9 (PHQ-9). The outcome of interest were the depressive symptoms as measured by PHQ-9 and will from now on be referred to as the outcome variable. Participants were randomly assigned to either an internet-based intervention (n = 97) or an attention-only waiting-list control condition (n = 91). Besides the outcome of interest, the dataset included other potential predictors such as gender and age. A summary of all variables part of this dataset can be found in appendix A.

In the original study of Van Luenen et al. (2018), the intent was to assess differences between the groups in changes in depressive symptoms from pretest to post-test by performing longitudinal multilevel regression analyses using the maximum likelihood estimation method. This can be seen as an appropriate strategy as multilevel models can handle missing data in the outcome variable using maximum likelihood estimation. It can be, however, less accurate than methods such as MI (Ibrahim & Molenbergh, 2009; Van Buuren, 2018). The missing data in this study again followed a monotone pattern and was collected at three time points. At the end of the study, for 57 of the 188 participants no information was gathered for at least one time point. Van Luenen et al. (2018) concluded that there were no differences in baseline characteristics between participants that dropped out and participants that completed the intervention. They claim that this indicated that none of the characteristics assessed for their study were associated with dropout and that therefore the results might be generalizable. This could be interpreted as data being MCAR, although in real life situations data being MCAR is highly unlikely and it is impossible to claim that missingness did not depend on the value of the outcome itself (Jelicic, Phelps & Lerner, 2009). They did not report about the relationship between baseline characteristics and dropout within each condition separately or at different timepoints. The influence of the outcome values that were observed was not mentioned either and therefore data being MAR seems more likely. Maximum likelihood estimation can be seen as an unbiased and efficient technique to account for missing data under the mechanism of MAR or MCAR.

4.2 Procedure

For this thesis, the results found by Van Luenen et al. (2018) will be replicated by fitting a multilevel regression analysis using maximum likelihood estimation on all available cases using lme() from the nlme package (Bates et al., 2021). As Van Leunen et al. (2018) did, time and group were included as fixed effects and slopes for time and the intercept as random effects. Here time was treated as a factor instead of a numerical value as in the simulation studies.

With the aim of investigating differences that arise in the estimated interaction effects from the missing data techniques, multiple imputation was used to generate complete datasets. On each of these complete datasets, the same multilevel regression analysis was performed and their results were pooled. When multiple imputation is used, the number of imputations has to be picked. More imputations will lead to more precise and replicable estimates. Von Hippel (2018) showed that the number of imputations needed depends on the amount of missing data to be imputed. Following his quadratic rule, 25 imputed datasets were chosen as optimal. The imputation model included the interaction between time and group, and all background variables (see appendix A) as predictors, as including as many variables as possible makes the assumption of MAR more likely (Schafer, 1997). Time was also included as a random and fixed effect in the imputation model in addition to the interaction between time and group. Multiple imputation was performed using the mice function from the package mice (Van Buuren & Groothuis-Oudshoorn, 2011). It is important to keep in mind that multiple imputation will only lead to unbiased estimates when the missing data mechanism is MAR.

Finally, missing data will also be imputed 25 times using the new technique of worst-case scenario imputation (R-function can be found in the appendix). Unlike the two techniques described above, the missing data mechanism is not assumed to be MCAR or MAR. worst-case scenario imputation is proposed to be an efficient technique to account for data that is missing under the mechanism of MNAR. More specifically, the missing data are assumed to represent the worst-case scenario. In this randomized controlled trial of Van Luenen et al. (2018) this means that dropout in the control condition is more likely for participants with the most progression, while dropout in the intervention condition is more likely for participants with a lack of progression concerning their depressive symptoms. WCSI will impute the missing values and capture this scenario. If data are MNAR and one does not account for this worst-case scenario, there is an increased chance that a Type I error will be made. WCSI uses the same imputation method as before with the exclusion of the

interaction as a predictor. All multilevel regression models have been fitted using the lme function from the nlme package (Pinheiro, Bates, DebRoy, Sarkar, 2021).

As the second simulation study showed that using higher percentages for WCSI resulted in the least biased estimates, worst-case scenario imputation will be used with only high percentages of the dataset (e.g. 75% and 100%). The performance of each missing data technique will be assessed by investigating differences in estimated effects that arise while keeping their underlying assumptions about the missing data mechanism in mind. The goal is not to find the best technique for this dataset but rather to discover the influence of making assumptions on model estimates.

4.3 Results

In this section, the results are discussed for each multilevel analysis using a different technique to handle missing data in the study of Van Luenen et al. (2018). All results are summarized in Table 7 - 10 and can be found on page 29 and 30.

Table 7 gives the results replicated from the study of Van Luenen et al. (2018) by fitting a multilevel regression analysis using maximum likelihood estimation on all available cases. When the original study is replicated, a significant negative interaction effect between time and group is observed, meaning that there are significant decreases of depressive symptoms in the intervention condition compared to control condition. This is observed from pretest to post-test 1 (β = -2.50, SE = .79, p = .002) and from pretest to post-test 2 (β = -2.07, SE = .86, p = .016). These replicated results are similar to the results found when multiple imputation was used, these can be seen in Table 8. Using multiple imputation replicates the significant differences between the intervention condition and control condition as seen by the negative interaction effects from pretest to post-test 1 (β = -2.37, SE = .79, p = .003) and pretest to post-test 2 (β = -1.96, SE = .87, p = .026). Although of less interest, these two techniques also result in similar estimated time effects. Both the results from ML and MI signal that the internet-based intervention is effective in reducing depressive symptoms in comparison to the attention-only waiting-list control condition.

Using worst-case scenario imputation, a technique that does assume that the missing data mechanism is MNAR, seems to alter the conclusions as the interaction effect becomes less strong. This can be seen from the pooled results in Table 9. When 75 percent of the information from the other condition is used for the imputation of missing values, the estimated interaction effects are still negative of sign but no longer found to be significant. This is observed both from pretest to post-test 1 (β = -1.38, SE = .79, p = .082) and pretest to

post-test 2 (β = -0.10, SE = .82, *p* = .901). This indicates that there are no longer any significant differences found between the control condition and intervention condition based on depressive symptoms over time when data are assumed to be MNAR. Using all information from the other condition for imputations, see Table 10, results is a weak but significant interaction effect between group and time observed from pretest to post-test 1 (β = -1.66 SE = .83, *p* = .046). The interaction effect between group and time from pretest to post-test 2 is not found to be significant (β = -1.18, SE = .89, *p* = .187), which contrasts with the significant results from ML (Table 5) and MI (Table 6).

Overall, it seems that assuming that missingness does not depend on the outcome itself (i.e. MAR or MCAR) in this specific dataset would lead to the conclusion that the internet-based intervention is effective in reducing depressive symptoms compared to the waiting-list control condition. This is true when both ML and MI were used. This conclusion would have been altered if data was assumed to be MNAR and missing data was imputed while taking the worst-case scenario into account. In this last case, the interaction effect from pretest to post-test 2 is no longer found to be significant and the interaction effect from pretest to post-test 1 is estimated to be weaker.

Table 7

Results from multilevel analysis using maximum likelihood estimation on all available cases

	Value	Standard error	DF	t-value	p-value
Intercept	11.110	0.472	279	23.548	<.001
Time (post 1) effect	-2.510	0.563	279	-4.460	<.001
Time (post 2) effect	-3.058	0.603	279	-5.070	<.001
Group effect	0.632	0.657	186	0.963	0.337
Time post 1 x Group effect	-2.495	0.789	279	-3.125	0.002
Time post 2 x Group effect	-2.074	0.859	279	-2.415	0.016

Table 8

Pooled results from multilevel analysis using maximum likelihood estimation on 25 datasets imputed using multiple imputation

	Value	Standard error	DF	t-value	p-value
Intercept	11.110	0.471	370	23.574	<.001
Time (post 1) effect	-2.605	0.580	212	-4.496	<.001
Time (post 2) effect	-2.946	0.581	225	-5.072	<.001
Group effect	0.632	0.656	370	0.964	0.336
Time post 1 x Group effect	-2.374	0.790	243	-3.004	0.003
Time post 2 x Group effect	-1.964	0.874	137	-2.246	0.026

Table 9

Pooled results from multilevel analysis using maximum likelihood estimation on 25 datasets imputed using worst-case scenario imputation using 75 percent of data

	Value	Standard error	DF	t-value	p-value
Intercept	11.110	0.471	370	23.574	<.001
Time (post 1) effect	-2.900	0.554	307	-5.234	<.001
Time (post 2) effect	-3.888	0.589	244	-6.604	<.001
Group effect	0.632	0.656	370	0.964	0.336
Time post 1 x Group effect	-1.381	0.792	261	-1.743	0.082
Time post 2 x Group effect	-0.103	0.824	235	-0.125	0.901

Table 10

Pooled results from multilevel analysis using maximum likelihood estimation on 25 datasets imputed using worst-case scenario imputation with 100 of data

	Value	Standard error	DF	t-value	p-value
Intercept	11.110	0.471	370	23.574	<.001
Time (post 1) effect	-2.825	0.585	241	-4.832	<.001
Time (post 2) effect	-3.348	0.617	195	-5.429	<.001
Group effect	0.632	0.656	370	0.964	0.336
Time post 1 x Group effect	-1.660	0.827	217	-2.008	0.046
Time post 2 x Group effect	-1.178	0.889	156	-1.325	0.187

Section 5. Discussion

This thesis introduced a new technique and performed the first investigation into its potential ability to handle missing not at random data in longitudinal randomized controlled trials. The performance of this technique is compared to maximum likelihood estimation on all available cases and multiple imputation using a Gibbs sampler for handling missing multilevel data by means of two simulations and one empirical study (Van Luenen et al., 2018).

The first simulation study showed that under the assumption of MAR, where missingness depended on the value of Y at the previous time point, maximum likelihood estimation performs best and results in only a small overestimation of the interaction effect. Multiple imputation showed very similar performance in this case and performed slightly better than maximum likelihood estimation under the assumption of MNAR, where missingness depended on the value of Y itself. In the first simulation, worst-case scenario imputation always performed worse than the other techniques. This was the case when data was manipulated to be both MAR and MNAR. This result was not surprising as WCSI assumes that the two groups have different reasons for dropout and that the missing values are representative of the worst-case scenario. Dropout in the first simulation study was generated to be similar for both groups and did not represent the worst-case scenario. As the amount of missing data increased from approximately 10 to 25 percent, all techniques showed less accurate performance, however maximum likelihood estimation and multiple imputation would still lead to unbiased estimates, especially when data was MAR. The first simulation study was important to include as it gave insight into the robustness of the WCSI procedure. It showed how strong the estimated effects were influenced by assuming the worst-case scenario when the missing data mechanism was in fact MAR or another form of MNAR.

The second simulation study focused on the situation where dropout indirectly depended on the outcome itself (i.e. true change in outcome) and did represent the worst-case scenario. This means that dropout in the control condition was simulated to be more likely for those that did show progression (e.g. placebo effect or natural recovery), while dropout was more likely in the intervention for those that did not show any progress or even adverse treatment effects. This simulation study first focused on datasets where an interaction effect was included. As the results have shown, maximum likelihood and multiple imputation on average result in overestimations of the interaction effect as indicated by the negative signs of their bias. ML and MI result in too strong estimated interaction effects, this might indicate that there is too much power as the probability that an effect will be detected is high. On the

other hand, WCSI resulted in underestimations of the interaction effect and could even result in estimated effects in the opposite direction. Although this means that conclusions are drawn with more caution, it can also be a sign that the estimated effects are too conservative. The results did indicate that the performance of WCSI increased as the amount of data used for imputation increased. Use of 100 percent of the data from the other condition for imputation resulted in unbiased estimates and high coverage probability for the datasets with approximately 10 percent missing data. Although the pattern described before was again observed in the datasets with approximately 25 percent missing data and worst-case scenario imputation with 100 percent was again the best performing technique, the estimated effects would still be biased to some degree. On the other hand, the coverage probability was very close to the preferred amount of 95 percent.

This second simulation study also investigated what would happen when the true interaction effect used to generate the data was set to zero (i.e. in the absence of an interaction effect). For the datasets with approximately 10 percent of missing data, using worst-case scenario imputation with any percentage always performed better than maximum likelihood estimation and multiple imputation based on all performance measurements (i.e. bias, root mean squared error and coverage probability). As before, ML and MI led to negative and stronger estimated interaction effects. Using WCSI with 10, 25 or 50 percent of the data resulted in smaller and positive estimated interaction effects. Using at least 75 percent of this for this technique would again give negative estimated interaction effects. In these simulations, using 75 percent of the data from the other condition for imputation of missing value can be seen as the best performing technique based on bias, RMSE and CP. The last part of the second simulation focused on results in the absence of a true effect and approximately 25 percent missing data. ML and MI again perform similarly and worse than WCSI as long as 25 percent of the data or more was used for imputation. Using all information from the other group (i.e. WCSI with 100 percent) resulted in negative bias as did ML and MI. Although all techniques used showed biased estimates, the effects were least biased when WCSI was used with 75 percent of the dataset.

The fourth section of this thesis includes a replication of the effects found using maximum likelihood estimation in the study of Van Luenen et al. (2018) and in addition the same effects were estimated using multiple imputation and worst-case scenario imputation. Maximum likelihood estimation and multiple imputation resulted in very similar estimated interaction effects. The conclusions made by Van Luenen et al. (2018) concerning effectiveness of the treatment would not have been altered. Those conclusions would have

been altered if WCSI was used. WCSI resulted in less strong estimated interaction effects and not all significant interaction effects were replicated. Results became more similar to the results found by Van Luenen et al. (2018) when a higher percentage of the data was used for imputation. The originally found significant interaction effect between group and time from pretest to post-test 2 did not appear with WCSI, even when 100 percent of the data was used for imputation.

In line with previous research (Enders & Bandalos, 2001; Schafer & Yucel, 2002; Black, Harel & McCoach, 2010; Cheema, 2014), maximum likelihood estimation and multiple imputation led to unbiased estimated effects when the data are missing at random, especially when only a small percentage of the data was missing. In the first simulation study, these techniques seemed to be applicable even to the simulation where missingness did depend on the outcome of Y itself. However, this will not always be the case and MI and ML can both lead to inaccurate estimates when data are MNAR as indicated by the second simulation study and previous research (Galimard et al. (2016). When data was manipulated to be MNAR and missingness depended on the true change, both MI and ML resulted in biased interaction effects when multilevel data was used to assess effectiveness of an intervention proposed to decrease symptoms.

As worst-case scenario imputation was developed as part of this thesis, there is no previous research available that can be used to (dis)confirm the results found. As expected, the new technique of WCSI performs better than MI and ML in the study where data are MNAR, representative of the worst-case scenario based on the true change of participants. Critics could argue that WCSI as a technique is only appropriate for this specific case of MNAR data and that its presence can never be proven with real data (Jelecic, Phelps & Lerner, 2009). This technique does however give more realistic estimated effects compared to sensitivity analysis and focuses on a situation where dropout inflates the probability of making a Type I error, therefore it stimulates more precaution when one makes claims about the effectiveness of a treatment. This was observed from the results using the dataset of Van Luenen et al. (2018), where both MI and ML resulted in similar effects that would lead to the conclusion that the intervention was effective for the reduction of depressive symptoms over time in comparison to the control condition. This was no longer the case when missing data was handled using worst-case scenario imputation, the interaction effect was no longer found to be significant from pretest to post-test 2 or even from pretest to post-test 1. These results, where there is a lack of treatment effectiveness, could be more accurate when the missing

data from the study by Van Luenen et al. (2018) turned out to be representative of the worst-case scenario.

Others could argue that the assumption of at least MAR in the dataset of Van Luenen et al. (2018) is plausible due to the inclusion of many predictors (Schafer, 1997). This does not mean that the presence of MNAR can be ruled out (Jelicic et al., 2009). To prove the presence of the worst-case scenario based on real data with missing values is impossible, but it does not seem to be a unique situation based on logical reasoning. It seems plausible that participants in a control condition (e.g. waiting-list) no longer want to wait for treatment when they already experience relief from their symptoms. On the other hand, it is not strange to think that participants in an intervention who do not experience any progression or even adverse effects decide to no longer participate in a study. When the worst-case scenario is assumed and WCSI is used for imputation, still finding a significant interaction effect can be interpreted as a stronger claim of a true effect. As mentioned before, maximum likelihood estimation and multiple imputation always resulted in estimated interaction effects that are stronger than the true effect in the second simulation study. These discrepancies between the estimated effects and the true effect grew larger as the amount of missing data increased and in the absence of a true interaction effect (Black, Harel & McCoach, 2010; Cheema, 2014). WCSI on the other hand always led to an underestimation of the true interaction effect or even caused an increase in Y (e.g. adverse effects). One could state that making assumptions based on logical reasoning is not enough and that it is best to focus on the observed data as it is all we have. However, the missing data can contain important information too and it is therefore of importance to emphasize that the worst-case scenario is also a situation that can easily result in finding a false positive effect. If dropout happens as proposed in this thesis but is handled as if it were MAR, this could have major consequences. It could lead to the implementation of an ineffective treatment in which time, money and effort are lost at the expense of the participants.

Although this was the first study into worst-case scenario imputation and more research is certainly needed, it seems that WCSI could be advised when MCAR nor MAR cannot be determined based on the data collected and the design of a randomized controlled trial makes the presence of a worst-case scenario plausible. For example, a placebo or waiting-list condition in comparison to a new developed intervention whose (potentially adverse) effects are unclear and not documented by previous research. The results of this thesis do encourage the use of maximum likelihood estimation and multiple imputation when data are assumed to be MAR (Enders & Bandalos, 2001; Schafer & Yucel, 2002; Newman,

2003; Grund, Ludtke & Robitzsch, 2016; Jakobsen & Gluud, 2017). Handling data that seems to be MAR using only worst-case scenario imputation is not advised. As can be seen from the first simulation study, the performance is worst for WCSI in this case. It is advised to include WCSI in addition to MI or ML to get an idea about what assuming MAR or MNAR (i.e. worst-case scenario) does to the results.

This thesis tried to emphasize that although it seems intuitive to make claims about missingness based on observed data as it is all we have, it can lead to results that are not representative of the truth. Missing data occurs in all longitudinal research and potentially contains information as interesting as the data observed. One should not disregard its influence easily or forget that data being MNAR can never be ruled out. Taking into account the potential influence of MNAR data might take more effort, it also provides more knowledge and certainty about the results found. This thesis has done a first investigation into the influence of data being MNAR and representative of the worst-case scenario on estimated effects and showed proof that WCSI can result in less biased estimated interaction effects in this scenario in comparison to ML and MI.

One of the strengths of the first simulation study is that it investigated the performance of different techniques under both MAR and MNAR with different percentages of missing data. The parameter estimates and dropout model were based on earlier research (Kenward & Diggle). For the second study, realistic parameter settings were selected based on the study by Van Luenen et al. (2018). This simulation study looked at performance with different amounts of missing data but also provided insight into what happened in the presence and absence of an interaction effect. Furthermore, using a real dataset has shown that making assumptions about the missing data can influence the conclusions about effectiveness of a treatment.

There were differences between the two simulation studies based on the (i) number of participants, (ii) number of time points, (iii) true parameters used for all effects, (iv) presence of a covariate, and most importantly, (v) dropout models. These factors were not varied within each simulation study to assess how it affects performance of each technique. In addition, the new technique was only compared in light of two well-established missing data techniques and was not compared to older techniques such as list-wise deletion or single mean imputation (Newman, 2003; Peugh & Enders, 2004) or more complex model-based approaches (Little, 1993; Little & Rubin, 2002). Both simulation studies and the real dataset had a monotone missing data pattern and missing data only occured in the dependent variable. Furthermore, only relatively small ($N < 200$) samples were considered. The

simulation studies can be best seen as “simplified versions of reality”, this makes the results less generalizable.

This thesis did also focus on performance of each technique using a real dataset where again the interaction effect between group and time was the point of interest, but it is hard to put the results from all studies together. One reason for this is that the empirical study included a decent amount of background variables and that the true missingness mechanism can never be known. Another reason is that the multilevel model on the true dataset was fitted by treating time as a factor variable, while time was treated as a numerical value in the simulation studies. This was done because it facilitated the specification of the data generation model. Normally, interactions between time and group used to assess group differences over time are estimated as in the empirical study (Van Luenen et al., 2018). Future research could focus more on simulation studies that more closely resemble true data by including more covariates or could even generate missing data in real datasets.

Overall, this thesis has shown that choosing a technique to handle missing data while assuming either MAR or MNAR can alter results to an extent where one would draw different conclusions. MI and ML are most appropriate when data are assumed to be MAR and datasets contain enough information to make that assumption more likely. If data being MNAR seems as, or even more, likely than MAR, for instance if there is not enough information available, it might be more appropriate to use WCSI. The use of WCSI can also be advised with caution when there are large amounts of missing data and one would like to have more certainty about the effectiveness of the study. For example, when logical reasoning could lead one to believe that the worst-case scenario is not unlikely or it is possible that the treatment has adverse effects. At this point in time, all well documented and available techniques assume MAR. Although these techniques, in particular maximum likelihood estimation, are easier to apply and require less additional effort than WCSI, convenience should never be a reason to assume MAR. Even when MAR seems most likely, results should always be interpreted with care and the possibility of MNAR must be kept in mind. Hopefully, with more research, WCSI can be used to handle MNAR data while accounting for the worst-case scenario in a more realistic way than, the often advised to use, sensitivity analyses. It is important to keep in mind that no matter what technique is used, missing data will always be a limitation when interpreting results. Missing data will always result in a loss of information, even if the data are MCAR.

References

Allison, P. D. (2011). *Missing data*. https://doi.org/10.4135/9781412985079.n8

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: implications of model misspecification. *Journal of Applied Statistics*, *38*(9), 1845–1865. https://doi.org/10.1080/02664763.2010.529882

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487–508. https://doi.org/10.3102/0034654314532697

Chen, N., Li, M., & Liu, H. (2018). Comparison of maximum likelihood approach, Diggle–Kenward selection model, pattern mixture model with MAR and MNAR dropout data. *Communications in Statistics - Simulation and Computation*, *49*(7), 1746–1767. https://doi.org/10.1080/03610918.2018.1506028

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, *43*(1), 49–73. https://doi.org/10.2307/2986113

Enders, C., & Bandalos, D. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/s15328007sem0803_5

Enders, C. K. (2011). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*(4), 267–288. https://doi.org/10.1037/a0025579

Fiero, M. H., Hsu, C. H., & Bell, M. L. (2017). A pattern-mixture model approach for handling missing continuous outcome data in longitudinal cluster randomized trials. *Statistics in Medicine*, *36*(26), 4094–4105. https://doi.org/10.1002/sim.7418

Galimard, J., Chevret, S., Protopopescu, C., & Resche‐Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, *35*(17), 2907–2920. https://doi.org/10.1002/sim.6902

Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. *Handbook of Psychology*. Published. https://doi.org/10.1002/0471264385.wei0204

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data. *SAGE Open*, *6*(4), 215824401666822. https://doi.org/10.1177/2158244016668220

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials - the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics & Gynaecology*, *125*(13), 1716. https://doi.org/10.1111/1471-0528.15199

Ibrahim, J. G., Chen, M. H., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*(2), 551–564. https://doi.org/10.1093/biomet/88.2.551

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, *18*(1), 1–43. https://doi.org/10.1007/s11749-009-0138-x

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1). https://doi.org/10.1186/s12874-017-0442-1

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(4), 1195–1199. https://doi.org/10.1037/a0015665

Kaciroti, N. A., & Raghunathan, T. (2014). Bayesian sensitivity analysis of incomplete data: bridging pattern-mixture and selection models. *Statistics in Medicine*, *33*(27), 4841–4857. https://doi.org/10.1002/sim.6302

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134. https://doi.org/10.2307/2290705

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ, United States: Wiley. https://doi.org/10.1002/9781119013563

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, *367*(14), 1355–1360. https://doi.org/10.1056/nejmsr1203730

Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., & Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, *21*(8), 1023–1041. https://doi.org/10.1002/sim.1064

Morris, T. P., Kahan, B. C., & White, I. R. (2014). Choosing sensitivity analyses for randomised trials: principles. *BMC Medical Research Methodology*, *14*(1). https://doi.org/10.1186/1471-2288-14-11

Musil, C. M., Warner, C. B., Yobas, P. K., & Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, *24*(7), 815–829. https://doi.org/10.1177/019394502762477004

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, *6*(3), 328–362. https://doi.org/10.1177/1094428103254673

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556. https://doi.org/10.3102/00346543074004525

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2015). The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0022-1

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*(3), 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. https://doi.org/10.1002/9780470316696

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data* (1st ed.). Abingdon, United Kingdom: Taylor & Francis. https://doi.org/10.1201/9781439821862

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989x.7.2.147

Schafer, J. L., & Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, *11*(2), 437–457. https://doi.org/10.1198/106186002760180608

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data Analysis*. Oxford, United Kingdom: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, *338*(jun29 1), 2393. https://doi.org/10.1136/bmj.b2393

van Buuren, S. (2018). *Flexible imputation of missing data* (Second edition). Retrieved from https://stefvanbuuren.name/fimd/

van Luenen, S., Garnefski, N., Spinhoven, P., & Kraaij, V. (2018). Guided internet-based intervention for people with HIV and depressive symptoms: a randomised controlled trial in the Netherlands. *The Lancet HIV*, *5*(9), 488–497. https://doi.org/10.1016/s2352-3018(18)30133-4

Verbeke, G., & Molenberghs, G. (2014). *Linear Mixed Models for Longitudinal Data*. New York, United States of America: Springer-Verlag. https://doi.org/10.1007/978-1-4419-0300-6

von Hippel, P. T. (2018). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, *49*(3), 699–718. https://doi.org/10.1177/0049124117747303

Wang, J., Xie, H., & Fisher, J. F. (2011). *Multilevel models: applications using SAS*. Berlin, Germany: De Gruyter. https://doi.org/10.1515/9783110267709

Zhu, X. (2014). Comparison of four methods for handling missing data in longitudinal data analysis through a simulation study. *Open Journal of Statistics*, *04*(11), 933–944. https://doi.org/10.4236/ojs.2014.411088

Appendix A. Descriptive statistics from Van Luenen et al. (2018)

Table A1

Descriptive statistics of the variables included in the empirical study (N = 188)

		M (SD) or %	Range
Gender			
	Female	88%	
	Male	12%	
Nationality			
	Dutch	84%	
	Other	10%	
	Dutch and other	6%	
Age		46.3 (10.6)	21 - 75
Relationship			
	Yes	54%	
	No	46%	
Education			
	Low	22%	
	Medium	41%	
	High	37%	
Employment			
	Yes	49%	
	No	51%	
Y (PHQ)		9 (5)	0 - 25

Appendix B. R-code worst-case scenario imputation

```
WCSI <- function(dataset, ntime = 3, m, perc, high = "3", low = "1", seed = 2021){
 #' @title: worst case scenario imputation for longitudinal data
 #' @description: uses Gibbs sampling procedure "2l.pan" from package "mice" to impute
 #' missing values in Y and returns m datasets
 #' This dataset needs to contain a variable to be imputed named Y.
 #' An id variable called "id" as a grouping variable.
 #' An sorted indicator of different timepoints called "time"(example 0,1,2,0,1,2,etc) .
 #' A grouping variable for condition called "group", where 0 signals control condition
 #' and 1 is indicative of intervention condition.
 #'
 #'
 #' @param: dataset is a dataset in long format to be used for imputations
 #' @param: ntime is the number of timepoints
 #' @param: m is the number of imputated data sets needed
 #' @param: perc is the percentage data from other conditions to use for imputations
 #' @param: high is the highest score to calculate change score
 #' @param: low is the lowest score to calculate change score
 #' @param: seed is the seed used for the multiple imputations
 #'
 #' @return: this function returns a list with three elements:
 #' datasets are the m imputed datasets
 #' originalimp_con is the result of imputations for the control condition
 #' originalimp_int is the result of imputations for the intervention condition
 #'
 #'
 #'
 # Load in the libraries needed for this function to work
 #library(dplyr)
library(mice)
library(pan)
library(miceadds)
 # Split data based on condition
 intervention_data <- subset(dataset, group == "1")
 control_data <- subset(dataset, group == "0")
 # get the change scores for intervention data
 # subset the scores at highest and lowest timepoints
 end_i <- subset(intervention_data, time == high)
 start_i <- subset(intervention_data, time == low)
 # calculate the change between timepoints
 changescore_i <- end_i$Y - start_i$Y
 # replicate change scores to be in line with length of data
 changes_i <- rep(changescore_i, each = ntime)
 # append the change scores to the data
 intervention_data$changescores <- changes_i

 # get the change scores for control data
 # subset the scores at highest and lowest timepoints
 end_c <- subset(control_data, time == high)
 start_c <- subset(control_data, time == low)
 # calculate the change between timepoints
 changescore_c <- end_c$Y - start_c$Y
 # replicate change scores to be in line with length of data
 changes_c <- rep(changescore_c, each = ntime)
 # append the change scores to the data
 control_data$changescores <- changes_c
```

```
# Index variable for complete cases of intervention data
ind.i <- complete.cases(intervention_data)
# Split into complete and missing
intervention_missing <- intervention_data[!ind.i, ]
intervention_missing <- intervention_data[intervention_data$id %in% intervention_missing$id, ]
intervention_complete <- intervention_data[!intervention_data$id %in% intervention_missing$id, ]

# Index variable for complete cases of control data
ind.c <- complete.cases(control_data)
# Split into complete and missing
control_missing <- control_data[!ind.c, ]
control_missing  <- control_data[control_data$id %in% control_missing$id, ]
control_complete <- control_data[!control_data$id %in% control_missing$id, ]

# For the complete intervention data
# Sort data based on change score and id
sorted_int <- intervention_complete[with(intervention_complete, order(changescores, id)), ]
# Calculate the number of row needed for highest percentage of changescores
top_perc <- ((nrow(intervention_complete)/100) * perc)
# Round number of rows needed to be a multiple of number of timepoints
top_perc <- round(top_perc / ntime) * ntime
# Select 1:top_perc rows to get participants with biggest change scores of intervention
best_int <- sorted_int[1:top_perc, ]
# Bind the missing of control with ... percentage best scoring of intervention
control_bestint <- rbind(best_int, control_missing)
# Sort data based on change score and id
sorted_cont <- control_complete[with(control_complete, order(-changescores, id)), ]
# Calculate the number of row needed for lowest percentage of change score
low_perc <- ((nrow(control_complete)/ 100 ) * perc)
# Round number of rows needed to be a multiple of number of timepoints
low_perc <- round(low_perc / ntime) * ntime
# Select 1:low_perc rows to get participants with smallest change scores of control
worst_cont <- sorted_cont[1:low_perc, ]

# Bind the missing of intervention with percentage of worst scoring control
intervention_worstcont <- rbind(worst_cont, intervention_missing)
# remove NA rows
library(dplyr)
NA.idx <- intervention_worstcont %>%
  is.na() %>%
  apply(MARGIN = 1, FUN = all)
intervention_worstcont <- intervention_worstcont[!NA.idx,]
# Bind with missing control and best intervention
control_bestint <- rbind(best_int, control_missing)
NA.idx <- control_bestint %>%
  is.na() %>%
  apply(MARGIN = 1, FUN = all)
control_bestint <- control_bestint[!NA.idx,]
# Remove best and worst from complete version
int_c <- intervention_complete[!row.names(intervention_complete) %in% row.names(best_int), ]
cont_c <- control_complete[!row.names(control_complete) %in% row.names(worst_cont), ]
# All complete data without missing values nor duplicated observations
observed_data <- rbind(int_c, cont_c)

# Delete changescores from dataframe by column number
change_num <- which(colnames(observed_data) == "changescores")
#delete <- match(removevar, names(observed_data))
observed_data <- observed_data[, -change_num]
```

```
# Set up imputations for intervention
intervention_worstcont <- intervention_worstcont[, -change_num]
# Make predictor matrix
pred_int <- make.predictorMatrix(intervention_worstcont)
pred_int["Y", "id"] <- (-2)
pred_int["Y", "time"] <- 2
pred_int["Y", "group"] <- 0
pred_int[, "group"] <- 0
# Specify methods
meth_int <- make.method(intervention_worstcont)
meth_int[1:length(meth_int)] <- ""
meth_int["Y"] <- "2l.pan"
# Create imputations
imp_int <- mice(intervention_worstcont, meth = meth_int, pred = pred_int, m = m,
            maxit = 1, print = FALSE)
# Get the datasets from "mids" object
datasets_int <- mids2datlist(imp_int)

# Set up imputations for control
control_bestint <- control_bestint[,  -change_num]
# Make predictor matrix
pred_con <- make.predictorMatrix(control_bestint)
pred_con["Y", "id"] <- (-2)
pred_con["Y", "time"] <- 2
pred_con["Y", "group"] <- 0
pred_con[, "group"] <- 0
# Specify method
meth_con <- make.method(control_bestint)
meth_con[1:length(meth_con)] <- ""
meth_con["Y"] <- "2l.pan"
# Generate imputations
imp_con <- mice(control_bestint, meth = meth_con, pred = pred_con, m = m, maxit = 1,
            print = FALSE)
# Get the datasets from "mids" object
datasets_cont <- mids2datlist(imp_con)

# Combine imputed datasets with observed
new_list <- list()
for( i in 1:m){
  new_list[[i]] <- rbind(data.frame(datasets_int[i]), data.frame(datasets_cont[i]),
                    observed_data)
}
output <- list()
output$datasets <- new_list
output$originalimp_con <- imp_con
output$originalimp_int <- imp_int
return(output)
}

###########
```